5 Project: Capstone Project 1: Data Wrangling (Supplier Pricing Prediction)

1. Create a Google Doc (1-2 pages) describing the data wrangling steps you took to clean the dataset. Include answers to these questions in your submission:

    1. **What kind of cleaning steps did you perform?**

    a) Data Conversion: To optimize computational resources, I changed data objects in to categorical values and dates in to date time objects.

    b) Fixed Quantity Mismatch: One of the attributes in the training dataset 'quantity' was wrongly assigned to cost/unit, causing error, hence fixed.

    c) Removed Duplicates: While building a combined data frame from 21 tables, some duplicated values had to be dropped from the final table.

    d) Applied Label Encoder: Final table had many categorical values (in the multiples of 2000+) which had to be converted using label encoder. I refrained from using one-hot-encoding to avoid 'curse of dimensionality'.

    e) Removed Infinity Number: One of the attribute 'component-id' in the secondary table had 'C-9999' as the model number, which didn't align with the rest of the other model numbers, hence had to be removed.

    2. **How did you deal with missing values, if any?**

    a) Missing Values 1St Treatment: Only 23% of the observations had complete information we could use to build the model. The rest 77% were null values that had to be rejected to simplify the merging steps.

    b) Missing Values 2nd Treatment: After merging the tables, more null values were rejected by working with observations that did not have null values.

    c) Missing Values 3rd Treatment: One of the attributes 'weight' had missing values, which were imputed using median, not mean to avoid outliers.

    d) Missing Values 4th Treatment: After merging the final table, I noticed that some tube assemblies had both null and null null values for ['weight', 'unique_feature', 'orientation', 'part_name'] and ['component_type_id', 'name'] which was filled using backward and forward filled methods.

    3. **Were there outliers, and how did you handle them?**

    a) Bend_Radius: Noticed that maximum bend radius for some of the assemblies were an infinity number '9999.00' resulting in an outlier. I checked the tube design handbook and learned that bend radius that is 7 times greater than the tube diameter are considered near flat, hence all greater values were replaced by 7 * Tube Diameter to limit peaks.

    b) Weight: Also, noticed some outliers in the tube's weight followed by increase in price which was left untreated, so that we could account for supplier pricing on heavy assemblies in our model.

----End---